基于模糊蚁群的加权蛋白质复合物识别算法 *

毛伊敏 [1]，刘银萍 [1]，胡　健 [2]

(1. 江西理工大学 信息工程学院，江西 赣州 341000; 2. 江西理工大学 应用科学学院 信息工程系，江西 赣州 341000)

摘　要：针对蚁群融合模糊 C-means (FCM)聚类算法在蛋白质相互作用网络中进行复合物识别的准确率不高、召回率较低以及时间性能不佳等问题进行了研究，提出一种基于模糊蚁群的加权蛋白质复合物识别算法 FAC-PC (algorithm for identifying weighted protein complexes based on fuzzy ant colony clustering)。首先，融合边聚集系数与基因共表达的皮尔逊相关系数构建加权网络；其次提出 EPS (essential protein selection)度量公式来选取关键蛋白质，遍历关键蛋白质的邻居节点，设计蛋白质适应度 PFC (protein fitness calculation)来获取关键组蛋白质，利用关键组蛋白质替换种子节点进行蚁群聚类，克服蚁群算法中因大量拾起放下和重复合并过滤操作而导致准确率和收敛速度过慢的缺陷；接着设计相似度 SI (similarity improvement)度量优化拾起放下概率来对节点进行蚁群聚类进而获得聚类数目；最后将关键蛋白质和通过蚁群聚类得到的聚类数目初始化 FCM 算法，设计隶属度更新策略来优化隶属度的更新，同时提出兼顾类内距和类间距的 FCM 迭代目标函数，最终利用改进的 FCM 完成复合物的识别。将 FAC-PC 算法应用在 DIP 数据上进行复合物的识别，实验结果表明 FAC-PC 算法的准确率和召回率较高，能够较准确地识别蛋白质复合物。

关键词：蛋白质相互作用网络；蚁群聚类算法；模糊 C-means；适应度；蛋白质复合物

中图分类号：TP399　　doi: 10.19734/j.issn.1001-3695.2018.10.0799

Algorithm for identifying weighted protein complexes based on fuzzy ant colony clustering

Mao Yimin [1], Liu Yinping [1], Hu Jian [2]

(1. *School of Information Engineering*, *Jiangxi University of Science & Technology*, *Ganzhou Jiangxi* 341000, *China*; 2. *Dept. of Information Engineering*, *College of Applied Science*, *Jiangxi University of Science & Technology*, *Ganzhou Jiangxi* 341000, *China*)

Abstract: Aiming at the problem that the accuracy and recall of the protein complexes identification algorithm based on ant colony and fuzzy C-means (FCM) clustering are not high and the running efficiency is low, this paper proposed a novel protein complex recognition algorithm named FAC-PC (algorithm for identifying weighted protein complexes based on fuzzy ant colony clustering) . Firstly, combing with the Pearson correlation coefficient and edge aggregation coefficient, it constructed the weighted protein network. Secondly, in order to overcome the defects of massive merger, filter, repeated pick-up and drop-down operations in ant colony clustering algorithm, it designed the EPS (essential protein selection) metric to select essential protein, and designed the PFC (protein fitness calculation) metric to traverse neighbors of essential proteins to obtain essential group proteins, then the essential group protein replaced the seed node in the process of ant colony clustering, which improved results that the accuracy and time performance. Furthermore, it proposed the SI (similarity improvement) metric to optimize the probability of picking and dropping operations of ant colony to obtain the number of clustering. Finally, according to the improved ant colony algorithm, it obtained the essential protein and the number of clustering to initialize the FCM algorithm, and designed the membership update strategy to optimize the membership update, at the same time, a new FCM objective function which took a balance between intra-clustering and proposed inter-clustering variation, finally identified the protein complex by improved FCM algorithm. It used FAC-PC algorithm to identify protein complexes on DIP data. The experimental results show that FAC-PC algorithm has better performance on accuracy and recall, which is more reasonable to identify protein complexes.

Key words: protein-protein interaction network; ant colony clustering algorithm; fuzzy C-means; fitness; protein complex

0　引言

蛋白质相互作用网络（protein-protein interaction，PPI）是指一个生命有机体内的所有蛋白质之间相互作用组成的网络，它可以表示成一个无向图[1]。在一个 PPI 网络中，蛋白质复合物是指在相同时间和空间通过相互作用组成一个多分子机制的蛋白质集合[2]。大量的生物实验和计算方法实验产生了许多高质量、大规模的 PPI 网络数据，这些数据为识别

收稿日期：2018-10-13；**修回日期：**2019-01-02　**基金项目：**国家自然科学基金资助项目（41562019，41530640）；江西省自然科学基金资助项目（GJJ161566）；江西省教育厅科技项目（GJJ151528GJJ181504）

作者简介：毛伊敏（1970-），女，江西赣州人，教授，博士，主要研究方向为数据挖掘、生物计算、地理信息系统（mymlyc@163.com）；刘银萍（1993-），女，硕士研究生，主要研究方向为数据挖掘、生物计算、地理信息系统；胡健（1967-），男，江西赣州人，教授，博士，主要研究方向为数据挖掘、软件工程.

蛋白质复合物奠定了基础，而蛋白质复合物的识别能够帮助人类预测未知的蛋白质功能，解释特定的生物进程，并为研究疾病的发生机理，寻找新的药物靶标，提供重要的理论基础[3]。因此，识别蛋白质复合物是生物信息领域中的一项研究热点。

迄今为止，利用计算方法进行蛋白质复合物识别已经是后基因组时代生物信息学领域中一个非常活跃的研究领域。根据计算机理的不同，识别蛋白质复合物的算法大体分为：基于密度的方法[4,5]、基于层次的方法[6,7]和基于划分的聚类方法[8,9]。这些方法都有一定的缺陷，基于密度的聚类方法很难对网络中大量的稀疏节点进行聚类，算法挖掘的功能模块的准确率不高；基于层次的聚类方法难于检测出节点交叠的功能模块，聚类结果对网络的噪声非常敏感。由于模糊 C-means (FCM) 聚类算法实现简单，收敛速度快和局部搜索能力强，利用模糊隶属度划分数据可以改进数据的硬划分问题。因此，目前 FCM 聚类算法已成功应用于 PPI 网络复合物识别，成为该领域的研究热点。Trivodaliev 等人[10]提出将 FCM 与谱聚类相结合用于蛋白质模块功能挖掘。该算法是根据数据节点的模糊隶属度将数据划分到不同的类中，实验划分结果却存在对初始聚类中心和聚类数目敏感的缺陷，隶属度矩阵更新较慢以及目标函数仅仅考虑类内之间的差异，没有考虑类间距对实验结果造成的影响，导致蛋白质复合物识别的过程容易陷入局部最优，算法的预测精度不高和收敛速度较慢。除此之外，近年来涌现出许多群智能思想融合图聚类过程的检测方法，该类算法通过模拟社会性生物群体间的协作行为实现复合物的检测挖掘，展现了良好的检测质量[11]。其中蚁群算法具有信息正反馈机制、并行性、全局化特征以及较强的鲁棒性特点，本身就可以直接聚类实现复合物的挖掘，因此基于蚁群的蛋白质复合物挖掘算法逐渐成为一新的研究热点。Ji 等人[12]提出蚁群聚类思想应用到 PPI 网络模块检测问题上，提出了基于蚁群聚类的 PPI 网络模块检测算法 ACC-FMD。赵学武等人[13]提出了融合时序保持特征和蚁群聚类的动态 PPI 网络复合物识别算法 ACC-DPC。这些算法的聚类过程存在反复拾起放下操作和大量的合并过滤操作，导致实验运行的时间效率以及准确性不高。为了克服 FCM 聚类算法对初始聚类中心和聚类数目敏感的问题，文献[14]提出将 FCM 与群智能人工蜂群聚类算法相结合用于蛋白质复合物识别，弥补 FCM 算法的不足，取得了较好的聚类效果。然而上述研究都是将 PPI 网络有效地用未加权图模型来描述，已经被证明可以比较有效地识别蛋白质复合物，但由于 PPI 网络本身的复杂性，可利用的 PPI 数据的不完整性以及 PPI 网络中存在噪声等众多问题，仅仅依靠 PPI 网络本身的蛋白质复合物研究已经受到了限制，实验结果容易受到假阳性以及噪声数据的影响；而且生物中每个蛋白质有着不同的功能，不同的边的重要性也不同，更真实、详尽地表达复杂蛋白质网络结构具有重要作用[15]。因此，将 PPI 网络构建为加权图来研究更为合理。目前从加权 PPI 网络图中识别蛋白质复合物越来越受到人们的关注。Dimitrakopoulos 等人[16]提出一种从加权蛋白质网络之中预测重叠蛋白质复合物的算法 GENA。Kouhsar 等人[17]提出一种快速高性能的 WCOACH 算法预测加权 PPI 网络中的蛋白质复合体。Ama 等人[18]提出一种跨模块中心移除的加权蛋白质复合物识别算法 IMHRC。这些方法克服了假阳性以及噪声对实验结果的影响，能很好地识别精度和很强的鲁棒性，但是聚类结果的召回率和时间效率不高。虽然基于加权 PPI 网络的复合物识别取得了一定的成效，但是如何有效地构建加权 PPI，如何克服复合物识别效果受假阳性的影响、蚁群聚类需大量的拾起放下和合并过滤操作，以及 FCM 算法对聚类中心和聚类数目敏感、隶属度更新较慢，目标函数仅仅考虑类内差异等导致的准确率、召回率不高以及执行效率低等缺陷，仍是亟待解决的问题。

针对以上问题，本文提出了基于模糊蚁群的加权蛋白质复合物识别算法 FAC-PC（algorithm for identifying weighted protein complexes based on fuzzy ant colony clustering），主要工作为：a）融合边聚集系数与基因共表达的皮尔逊相关系数构建加权蛋白质网络；b）设计基于 PPI 网络拓扑特性与基因表达数据的 EPS 度量公式选取关键蛋白质；c）提出基于期望稠密度和模块度的 PFC 度量公式获取关键组蛋白质；d）设计基于权重的相似度 SI 度量优化蚁群算法的拾起放下概率，完成蚁群聚类获得聚类数目；e）利用蚁群聚类获得的聚类数目和关键蛋白质初始化 FCM 聚类算法，设计隶属度更新策略来改进 FCM 隶属度的更新计算，同时综合考虑类内和类间距，改进 FCM 算法的目标函数， 最后利用改进的 FCM 完成复合物的识别。实验结果表明本文算法运行效率高，聚类结果的准确率以及召回率较高。

1 相关工作

1.1 FCM 聚类算法

FCM 聚类算法[19]通过计算每个样本点对所有类中心的隶属度，并对目标函数不断进行优化找到最优解，从而决定样本点的隶属，达到对样本数据集进行聚类的目的。

设数据集 $X=\{X_1,X_2,X_3,...,X_n\}$，其中 $X_i=\{X_{i1},X_{i2},...,X_{in}\}$。

FCM 算法在满足约束条件 $\sum_{j=1}^{C}u_{ij}=1,i=1,2,...N$ 的前提下最小化目标函数 J，目标函数 J 定义如下：

$$J(u,c,k)=\sum_{i=1}^{N}\sum_{j=1}^{K}u_{ij}{}^{m}d(x_i,c_j) \tag{1}$$

其中：$m>1$ 为模糊加权指数；类别数为 K，$d(x_i,c_j)$ 为 x_i 与 c_j 间的欧式距离。结合拉格朗日最小二乘法原理，最小化目标函数得到隶属度 u_{ij} 和聚类中心 c_j 的迭代更新表达式如下：

$$u_{ij}=\left[\sum_{k=1}^{C}\left(\frac{d(x_i,c_j)}{d(x_i,c_k)}\right)^{\frac{1}{m-1}}\right]^{-1} \quad i=1,2,...,N;j=1,2,...,K \tag{2}$$

$$c_j=\frac{\sum_{i=1}^{N}u_{ij}^{m}\cdot x_i}{\sum_{i=1}^{N}u_{ij}^{m}},j=1,2,...,K \tag{3}$$

该算法本质上是一种局部搜索寻优方法，计算简单，容易实现，但对初始聚类中心极为敏感，容易陷入局部极值而很难得到全局最优解，聚类个数需要人为设定，隶属度更新较慢以及目标函数仅仅考虑类内距，没有考虑类间距，这给蛋白质复合物挖掘造成十分不利的影响，因此采用传统的 FCM 算法无法对蛋白质复合物进行准确的挖掘。

1.2 蚁群聚类算法

本文采用 ACC-FMD 算法[12]的蚁群聚类思想来介绍蚁群聚类过程。该方法将 PPI 网络看成是无向图 $G(V,E)$，其中：V 表示蛋白质节点集合；E 表示蛋白质相互作用边的集合。主要过程为：

a）选取种子节点。根据节点的聚集系数，将聚集系数大于阈值的节点选择出来。对于网络 $G(V,E)$ 中的任意节点 i,其聚集系数定义为

$$\phi_i = \frac{2n_i}{\{Neigh(i)\|Neigh(i)-1|}\tag{4}$$

其中：$Neigh(i)$ 表示节点 i 的直接邻居集合；n_i 表示集合 $Neigh(i)$ 中的节点之间的相互作用的个数。

b）聚类过程。蚂蚁开始遍历种子节点的领域，根据拾起放下概率模型来完成聚类。固拾起放下概率模型定义如下：

$$P_p(j)=\left[\frac{k_p}{k_p+s(i,j)}\right]^2\tag{5}$$

$$P_d(j)=\begin{cases}2s(i.j) & s(i,j)<k_d\\ 1 & 其他\end{cases}\tag{6}$$

其中：$s(i,j)$ 是节点 i 和 j 的结构相似性；k_p 和 k_d 两个参数。结构相似度定义为

$$s(i,j)=\frac{|\tau(i)\cap\tau(j)|}{\sqrt{|\tau(i)\cap\tau(j)|}}\tag{7}$$

其中：$\tau(i)$ 是由节点 i 和其直接邻居节点构成的集合。

c）信息传递。通过节点之间的相似性将上一代的最优解传递给下一代。该算法评价解的质量是通过模块化密度来实现的，每一代根据 D 值选取最优解，其定义为

$$D=\sum_{h=1}^{m}\frac{2l_h-\overline{l_h}}{n_h}\tag{8}$$

其中：m 是预测得到的复合物的数量；l_h 是复合物 h 中的边数；$\overline{l_h}$ 为边的一端在复合物内部，另一端在复合物的外部的边的数量；n_h 是复合物中节点的个数。

d）后处理过程。对初始聚类结果进行合并和过滤两个基本后处理操作。合并操作是合并两个相似度大于阈值的模块，对合并后的聚类结果过滤掉那些密度小于阈值的模块，其相似度定义为

$$S\left(M_x,M_y\right)=\frac{\sum_{i\in M_x,j\in M_y} r(i,j)}{\min\left\{|M_x|,|M_y|\right\}}\tag{9}$$

$$r(i,j)=\begin{cases}1 & i=j\\ \frac{|g_i\cap g_j|}{|g_i\cup g_j|} & i\neq j,(i,j)\in E\\ 0 & 其他\end{cases}\tag{10}$$

蚁群聚类是一种启发式的搜索算法，具有正反馈性、自组织性和健壮性等优点，聚类过程却存在大量的重复合并过滤以及拾起放下操作，导致聚类的时间效率以及准确率较低，因此采用传统的蚁群算法无法有效地对蛋白质复合物进行准确挖掘。

2 FAC-PC 算法

2.1 算法思想

针对蛋白质相互作用网络存在不稳定性，复合物的识别效果容易受到假阳性的影响；蚁群聚类效率受合并过滤以及重复拾起放下操作影响较大和 FCM 聚类结果对初始聚类中心、聚类数目敏感、隶属度更新较慢以及目标函数仅仅考虑类内之间的差异等问题，为提高算法的准确率、召回率、执行效率和降低假阳性的影响，本文借鉴文献[14]的群智能算法融合 FCM 算法来实现蛋白质复合物的挖掘思想，利用蚁群算法的信息正反馈机制、并行性、全局化特征以及较强的鲁棒性特点来解决 FCM 对聚类中心和聚类数目敏感的问题，提出了一种在加权 PPI 网络中有效地识别蛋白质复合物算法 FAC-PC。具体 FAC-PC 算法思想为：首先以蛋白质相互作用网络为框架，利用边聚集系数与基因共表达的皮尔逊相关系数衡量相互作用边的可靠性进而构建加权网络；其次利用 EPS 度量公式选取关键蛋白质，遍历关键蛋白质的邻居节点，利用本文设计的 PFC 度 量来获取关键组蛋白质，将关键组蛋白质替换种子节点进行蚁群聚类；接着使用相似度 SI 度量优化蚁群拾起放下概率来对节点进行蚁群聚类，获得聚类数目；最后将关键蛋白质和通过蚁群聚类得到的聚类数目初始化 FCM 算法，利用改进的隶属度更新策略来优化隶属度的更新，同时也提出兼顾类内距和类间距的 FCM 迭代目标函数，并利用改进的 FCM 算法最终完成复合物的识别。

2.2 加权蛋白质网络的构建

针对传统的基于 PPI 网络的蛋白质复合物识别算法的准确度比较依赖于网络本身的可靠性，复合物的识别效果容易受到假阳性以及噪声数据的影响；生物蛋白质网络中不同边的重要性不同等问题，本文基于蛋白质复合物是成簇出现且倾向于共表达的事实，且蛋白质复合物的挖掘与相互作用的可靠程度之间关系密切，综合考虑蛋白质网络的拓扑边聚集系数和共表达生物特征，采用边聚集系数[20]和皮尔逊相关系数[21]衡量蛋白质两个节点之间相互作用的可靠程度。

边聚集系数作为蛋白质相互作用网络的一个重要拓扑特性，可以用来描述蛋白质相互作用之间的可靠性，还可以用来衡量蛋白质之间属于用一簇的概率，能够较好地识别出关键蛋白质。则蛋白质节点 u 和 v 的边聚集系数 $ECC(u,v)$ 计算如下：

$$ECC=\frac{|N(u)\cap N(v)|}{\min(N_u-1,N_v-1)}\tag{11}$$

其中：$|N(u)\cap N(v)|$ 表示节点 u 和 v 所共有的邻居节点数；N_u 和 N_v 分别表示节点 u 和 v 的度。

另一方面利用基因表达数据来计算两个蛋白质节点之间的皮尔逊相关系数，以衡量蛋白质之间的可信度。$PCC(u,v)$ 表示节点之间的皮尔逊相关系数，则节点 u 与 v 之间的皮尔逊相关系数计算如下：

$$PCC(u,v)=\frac{1}{k-1}\sum_{t=1}^{k}\left(\frac{E_{xp}(u,t)-\overline{E_{xp}(u)}}{\sigma(u)}\right)\left(\frac{E_{xp}(v,t)-\overline{E_{xp}(v)}}{\sigma(v)}\right)\tag{12}$$

其中：N_u 和 N_v 分别代表节点 u 和 v 的直接邻居集合；k 为样本数；基因表达数据中时刻数 $E_{xp}(u,i)$ 和 $E_{xp}(v,i)$ 分别为蛋白质 u 和 v 在时刻 i 下的表达值；$\overline{E_{xp}(u)}$ 和 $\overline{E_{xp}(v)}$ 为蛋白质 u 和 v 在所有时刻下的平均表达值；$\sigma(u)$ 和 $\sigma(v)$ 表示蛋白质节点 u 和 v 在所有时刻下的标准方差。

在 PPI 网络拓扑边聚集系数的基础上，融合基因共表达数据，设计出了边聚集系数与基因共表达的皮尔逊相关系数的乘积公式用于计算相互作用边的存在概率，从而构建加权蛋白质相互作用网络。则蛋白质相互作用网络中边的权重 $P(u,v)$计算如下：

$$P(u,v)=ECC(u,v)\times PCC(u,v)\tag{13}$$

通过式（13）构造的加权网络，不仅考虑了节点的拓扑特性聚集程度，而且还增加了皮尔逊相关系数来度量相互作用蛋白质的基因共表达强弱程度，可以将一部分权值为 0 的数据排除，降低预测方法对蛋白质相互作用网络本身可靠性的依赖程度以及假阳性和噪声数据对实验造成的影响，进而提高识别的准确度。加权后的蛋白质网络形式化定义如下：

定义 1 加权蛋白质网络 $DG(V,E,P)$。其中：$V=(v_1,v_2,v_3,v_4,v_4,\ldots,v_n)$ 表示蛋白质节点集合；$E=(e_1,e_2,e_3,e_4,e_5,\ldots e_m)$ 表示蛋白质相互作用的集合；

$P=(p(e_1),p(e_2),p(e_3),p(e_4),p(e_5),...p(e_m))$ 表示相互作用权重的集合。

2.3 蚁群算法的改进

2.3.1 关键蛋白质的选取

针对蚁群聚类种子节点的选取仅仅依靠 PPI 数据，识别的准确率比较依赖于网络本身的可靠性，实验结果容易受到假阳性的影响。为了降低假阳性的影响，提高聚类准确性，本文设计出基于边聚集系数和皮尔逊相关系数的关键蛋白质选取 EPS 度量公式。

给定蛋白质加权网络 $DG(V,E,P)$， $ECC(u,v)$ 表示节点 u 与 v 之间的边聚集系数， $PCC(u,v)$ 表示节点之间的皮尔逊相关系数，则关键蛋白质选取 EPS 度量公式为

$$EPS(u)=\delta+\sum_{(u,v)\in E}\frac{P(u,v)-\min P(u,v)}{\max P(u,v)-\min P(u,v)} \tag{14}$$

$EPS(u)$考虑到了节点 u 和 v 在网络的拓扑特性的聚集程度，还增加了基因共表达程度来衡量一个节点和其邻居节点成簇的可能性，而且还考虑到基因表达数据与 PPI 网络数据的差别，因此能有效地评价一个蛋白质的关键性。根据式（14），将高于关键性阈值的节点选取出来，这样可以降低假阳性和假阴性对实验结果产生的影响，而且使得有公共顶点的稠密子图的相似度尽可能降低，最终提高算法运行的准确率，本文设置 δ=0.01 。

关键蛋白质的选取思想如下：首先基于蛋白质复合物是成簇出现且倾向于共表达的事实，利用式（13）来计算边的存在概率,构建加权网络；接着充分考虑节点之间的紧密程度以及共表达程度，利用 EPS 度量式（14）来计算网络节点的权重，将节点权重大于关键性阈值的节点选取出来作为关键蛋白质。这是因为关键性高的蛋白质对生命活动更为重要，从而求得的蛋白质复合物更能体现功能模块的生物特性，而且相比非关键蛋白质，关键蛋白质对于复合物的挖掘的重要性更高。

关键蛋白质的选取过程形式化如下：

输入：蛋白质网络 $G(V,E,P)$ ，关键性阈值 θ ，基因表达数据和参数 δ 。

输出：关键蛋白质集合 $\{v_1,v_2,...,v_k\}$ 。

a)构建加权蛋白质网络

(1) for each $(u,v)\in E$ do

(2) Compute $ECC(u,v)$ by Eq. (11)

(3) Compute $PCC(u,v)$ by Eq. (12)

(4) Compute $P(u,v)$ by Eq. (13)

(5)end for

b)选取关键蛋白质

(1) $L=\varnothing$

(1) for each $v_i\in V$ do

(2) If $EPS(v_i)>\theta$ do

(3) $L=L\cup v_i$ 将 L 中的关键蛋白质按照 EPS 权重非递减排序, $L=\{v_1,v_2,...,v_k\}$ 且 $EPS(v_1)\geq EPS(v_2)\geq ...EPS(v_k)$

(4) end if

(5)end for

2.3.2 关键组蛋白质的形成

针对蚁群聚类采用种子节点来扩展形成蛋白质复合物，若两个种子节点之间相似度比较大，那么聚类形成的两个复合物也比较相似，在后处理过程中，需要将这两个模块合并，而合并操作需要大量的操作计算，影响算法的时间性能；而且一个蛋白质节点可能处于多个种子节点的邻域内，蚁群算法需重复拾起放下操作导致的算法运行效率不高和准确率低等问题。为了提高实验运行效率以及识别出高内聚低耦合的复合物，本文提出蛋白质适应度 PFC 度量来选取关键组蛋白质，进而利用关键组蛋白质替代种子节点进行蚁群聚类。这是因为关键组蛋白质是一个子图，由于子图之间的差异性比原来蚁群聚类算法中种子节点之间的差异性大，所以求得的复合物不需要再进行合并，从而提高了算法的运行效率以及准确率；同时在每次拾起放下之前会判断加入该蛋白质节点和不包含该蛋白质的适应度差值是否大于 0，如果适应度小于 0，不对它进行拾起放下操作，减少拾起放下次数进而提高计算效率。

基于复合物是稠密子图且具有高度的模块性，设计出基于期望稠密度[22]和模块度[23]的蛋白质节点适应度 PFC 度量公式，根据节点适应度来遍历关键节点的邻域节点，最终形成关键组蛋白质，利用核心组蛋白质替代种子节点进行蚁群聚类。

给定一个蛋白质网络 $DG(V,E,P)$，其中 $ED=\frac{2\times\sum_{i=1}^{m}p(e_i)}{|V|\times(|V|-1)}$ ，$E=(e_1,e_2,e_3,e_4,e_5,...e_m)$ ，和子图 S，则网络图 G 的期望稠密度定义为

$$ED=\frac{2\times\sum_{i=1}^{m}p(e_i)}{|V|\times(|V|-1)} \tag{15}$$

模块度的定义为

$$WR=\frac{\sum_{v\in S}p(v_a,S)}{\sum_{v\in S}p_{out}(v_a,S)+\sum_{v\in S}p(v_a,S)} \tag{16}$$

其中：P 表示网络图中边的存在概率，通过边聚集系数和皮尔逊相关系数的乘积计算得到； v_a 是子图 S 中的任意节点；$P(v_a,S)$ 是节点 v_a 与子图 S 内部其他节点的连接边的权重；$P_{out}(v_a,S)$ 是节点 v_a 与 DG-S 中其他节点的连接边的权重。该公式充分考虑到内部节点与外部节点之间的联系。考虑到关键组蛋白质是小而稠密的模块子图，为了挖掘高内聚低耦合的关键组蛋白质，进而提高蛋白质复合物识别的准确性，本文提出的子图适应度以及蛋白质适应度 PFC 的计算公式分别如下：

$$F_S=\rho ED\times(1-\rho)WR \tag{17}$$

$$PFC(S,v)=F_{S+v}-F_{S-v} \tag{18}$$

其中： $\rho\in(0,1)$ 表示期望稠密度和模块度这两种特征在 F_S 所占的重要程度， ρ 越大，说明子图密度在 F_S 中影响力越大，而子图模块度的影响力越小； $S+v$ 与 S-v 分别表示在子图 S 加入节点 v 和删除节点 v 的聚类； $PFC(S,v)$ 表示为子图 S 含节点 v 和不包含节点 v 时的节点适应度之间的差值，当 $PFC(S,v)$ 越大，则节点 v 越可能属于子图 S。

关键组蛋白质的形成思想如下：首先基于加权蛋白质网络，遍历关键蛋白质节点的邻域节点，把关键蛋白质节点当作初始的关键组蛋白质子图 Set；然后根据式（18）来判断是否将关键蛋白质的邻域节点添加进来，若节点 v 使得 $PFC(Set,v)$=$F_{Set+v}-F_{Set-v}>0$ ，则添加到关键组子图中，逐渐遍历关键节点的邻域，直到所有邻域节点都被遍历完或这些顶点对子图的适应度都为负，扩展的过程将结束，得到核心组蛋白质 Set。

关键组蛋白质的形成过程如下：

输入：蛋白质网络 $DG(V,E,P)$，关键蛋白质 $L=\{v_1,v_2,...,v_k\}$ 。

输出：关键组蛋白质集合 Set。

(1) $Set=\varnothing$

(2) For i=1 to $|L|$ do

(3) $Set_i = Set_i \cup L(v_i)$

(4) For $v_j \in Neigh(L(v_i))$ do//遍历关键蛋白质的邻域节点

(5) If $PFC(Set_i, v_j) > 0$ do

(6) $Set_i = Set_i \cup v_j$ // $Set \leftarrow v_j$ 输出关键组蛋白质集合 Set

(7) End if

(8) End for

(9)End for

例如，图 1 给出了一个包含节点 v_2 v_3 v_4 的加权网络图。根据节点适应度来判断是否添加 v_1，若添加节点使得适应度差值大于 0，则添加节点，否则不添加节点。具体节点适应度的计算过程如下：

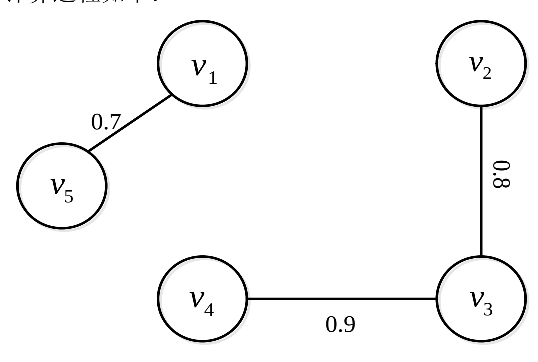


图 1　实例图

Fig. 1　Instance graph

由图 1 得知，$p(v_2,v_3)=0.8$，$p(v_3,v_4)=0.9$ $p(v_1, v_5)=0.7$，若节点 v_1 添加到该图中，得到 $p(v_1,v_2)=0.8$ $p(v_1,v_4)=0.7$，根据式（18）来计算添加节点 v_1 和删除节点 v_1 的节点适应度的差值，若大于 0，则添加否则不添加。

$$F_{S+v} = 0.1\times ED\times 0.9\times WR = \frac{0.1\times 2\times(0.8+0.8+0.9+0.7)}{4\times 3}\times 0.9\times\frac{0.8+0.8+0.9+0.7}{0.8+0.9+0.8+0.8+0.9+0.7} = 0.0313$$

$$F_{S-v} = 0.1\times ED\times 0.9\times WR = \frac{0.1\times 2\times(0.8+0.9)}{2\times 3}\times 0.9\times\frac{0.8+0.9}{0.7+0.8+0.9} = 0.0361$$

因为 $F(S,v)$=$F_{S+v} - F_{S-v} < 0$，所以不添加节点 v_1。

2.3.3 相似度改进的 SI 度量

针对蚁群算法需要反复计算节点的邻居节点数进而归一化共同邻居节点，造成算法的运行效率不高的问题，本文根据节点与核心组蛋白质的相似度 SI 度量公式来计算拾起放下的概率，再利用该模型完成聚类。下面给出相似度的 SI 度量计算公式。

给定无向图 $DG(V,E,P)$,其中：V 表示蛋白质节点集合；E 表示相互作用的集合；$P(i,j)$ 为蛋白质节点 i 与 j 通过边聚集系数以及基因共表达信息乘积进行加权得到的边权重，则蛋白质 i 与关键组蛋白质 Set 的相似度 SI 度量公式如下所示：

$$SI(i,Set) = \frac{\sum_{j\in Set} P(i,j)}{|Set|(|Set|-1)/2} \tag{19}$$

证明：

a) 对于 $\forall i, Set$，$SI(i,Set) = s(Set,i)$，对称性满足；

b) 对于 $\forall i, Set$，$\sum_{j\in Set} P(i,j) \ge 0$ 且 $|Set|(|Set|-1)/2 > 0$ 则 $SI(i,Set) \ge 0$，非负性满足；

c) 对于 $\forall i,z,Set$，$SI(i,z)$+SI$(z,Set) \ge SI(i,Set)$，三角不等式满足。

因此式（19）满足相似度度量的对称性、非负性、和三角不等式特性，且满足全局一致性聚类假设和局部一致性假设，是相似度度量公式。

本文在聚类的过程中，根据节点与关键组蛋白质的相似度 SI 度量公式来计算拾起放下的概率，再利用该模型完成聚类，故将式（5）和（6）的拾起放下概率模型公式转变为

$$P_p(i) = \left[\frac{k_p}{k_p + SI(i,Set)}\right]^2 \tag{20}$$

$$P_d(i) = \begin{cases} 2SI(i,\text{Set}) & ,SI(i,\text{Set}) < k_d \\ 1 & 其他 \end{cases} \tag{21}$$

基于改进的蚁群算法，具体获得蛋白质复合物的聚类数目思想如下：根据 2.3.1 节选取出来的关键蛋白质作为初始的关键组蛋白质 Set，遍历关键蛋白质的邻域节点，然后根据式（18）来判断是否将关键蛋白质的邻域节点添加进来，若节点 v 使得 $PFC(Set,v)$=$F_{Set+v} - F_{Set-v} > 0$，则添加到关键组子图中，再根据 SI 度量来计算节点与关键组蛋白质子图之间的相似度，根据拾起放下概率模型来实现蚁群聚类，重复迭代获得最优解，输出蚁群聚类的聚类数目。蚁群聚类的具体过程如下：

输入：蛋白质网络 $DG(V,E,P)$，关键蛋白质 $L=\{v_1,v_2,...,v_k\}$。

输出：蛋白质复合物的聚类数目 M。

(1)M=0, $Set=\varnothing$

(2)while $L\neq\varnothing$ do

(3) for t=1 to I do

(4) $Set_i = Set_i \cup L(v_i)$

(5) for n=1 to N do

(6) for i=1 to $|L|$ do

(7) 调用关键组蛋白质的形成过程，得到集合 Set

(8) for $v_x \in Neigh(Set_i)$ do

(9) if $PFC(Set_i, v_x) > 0$ do

(10) 利用拾起放下概率模型对节点进行聚类

(11) end if

(12) end for

(13) 蚂蚁 n 得到自身的解

(14) end for

(15) 第 t 代蚂蚁全部得到自身的解，根据模块度评价标准，求出本次迭代的最优聚类结果

(16) M=M+1

(17) end for

(18) end for

(19)Return M //输出蛋白质复合物的数目

本文将关键蛋白质节点以及通过改进的蚁群聚类算法获得的聚类个数初始化 FCM 算法，弥补 FCM 算法对初始聚类中心和聚类数目敏感的缺陷。

2.4 FCM 算法的改进

2.4.1 隶属度更新的改进策略

针对 FCM 算法的聚类实际上就是一个隶属度矩阵 u 和聚类中心 c 交替优化过程，当隶属度较大时，蛋白质节点所属的类别不发生改变以及隶属度更新较慢等问题，为了提高算法的收敛速度，可以修正隶属度矩阵 u 来计算下一次迭代的聚类中心，使计算结果更加合理，提高算法的收敛速度。隶属度越大，样本对类中心的吸引力越大。本文基于竞争学习的思想，给出一种隶属度的改进策略：在通过蚁群算法得到初始的聚类中心和聚类数目之后，得到较为可靠的隶属度值，使得距离样本中心点的类中心作为获胜节点，距离次者

的节点作对手，通过加入一个抑制参数来不同幅度减弱对手来加快赢着的收敛速度，进而提高算法的执行效率。具体描述为：对于对象 x_i ,若它对第 t 类的隶属度最大，为 u_{ti}，与同一行的剩余隶属度相差较大时，整体的隶属度的更新速度加快的较多；对第 s 类的隶属度为次大，为 u_{si}，给定一个抑制参数 η，则隶属度的更新公式为

$$u_{ti}^{'}=u_{ti}+\eta u_{si},\ u_{si}^{'}=(1-\eta)u_{si} \tag{22}$$

由式(2)可以看出，在隶属度更新时，本文充分考虑及节点自身的状态，若对象 x_i 对 t 类的隶属度以及对 s 类的隶属度相差的不大，那么上述公式的更新速度就会变慢；若有明显的优势时，即加快隶属度的更新速度。公式中 η 的取值会直接影响到算法的执行效率，本文通过实验验证，将参数设置为 η=0.6 。

2.4.2 FCM 目标函数选取的改进

针对传统的 FCM 算法的目标函数仅仅考虑了类内距离，没有考虑类间距，采用梯度法求解极值，所求解容易陷入局部最优，导致复合物挖掘的准确度不高。为了避免算法陷入局部最优，挖掘出高内聚低耦合的复合物，综合考虑类间距和类内距，本文根据 Xie-Beni 提出的聚类有效性指标[24]，给出一种兼顾类内和类间距的新 FCM 的目标函数。

类内距差异 $W(u,c,k)$ 和类间距 $A(u,c,k)$ 差异分别为

$$W(u,c,M)=\sum_{i=1}^{|V|}\sum_{j=1}^{M}u_{ij}^{m}d(x_i,c_j) \tag{23}$$

$$A(u,c,M)=|V|\times\min_{j\neq M}\|c_j-c_M\|^2 \tag{24}$$

综合考虑类内和类间距差异，改进 FCM 算法的目标函数，即

$$J(u,c,M)=\frac{W(u,c,M)}{MA(u,c,M)}=\frac{\sum_{i=1}^{|V|}\sum_{j=1}^{M}u_{ij}^{m}d(x_i,c_j)}{M|V|\min_{j\neq M}\|c_j-c_M\|^2} \tag{25}$$

2.5 算法描述

FAC-PC 算法具体的实现步骤如下：

a) 初始化参数设计。蚂蚁数目为 N，蚁群迭代次数 I，拾起参数 k_p，放下参数 k_d，关键阈值 θ、ρ、δ、ω、ε、η。

b) 根据式（13）对蛋白质网络边进行加权，进而采取式（14）来计算 $v_i\in V$ 节点的 EPS 值。若，$EPS(v_i)>\theta$ 将该节点加入到关键蛋白质集合 L,按照非递减的顺序进行排序，将权值 $P(u,v)$=0 的边当做噪声数据移除。

c) 将关键蛋白质节点当做初始的关键组子图 Set，遍历关键节点的邻域节点，根据式（18）计算节点 v 的适应度，若节点 v 使得 $PFC(Set,v)=F_{Set+v}-F_{Set-v}>0$，则添加到关键组子图 Set 中。重复操作，直到所有邻域节点都被遍历完或这些顶点对子图的适应度都为负，扩展的过程将结束，得到关键组蛋白质 Set。

d) 根据式（18）计算 v 对于关键组蛋白质 Set 的节点适应度，若 $PFC(v,Set)>0$，利用式（20）和（21）计算拾起放下概率，利用拾起放下规则来完成蚁群聚类过程；否则不进行操作，直接转向步骤 f)。

e) 重复步骤 d)，根据改进的蚁群算法得到聚类个数 M。

f) 将上述得到的关键蛋白质节点集合 L 和聚类个数 M 初始化 FCM 聚类算法的初始聚类中心 c 和聚类个数 M，根据式（2）计算隶属度矩阵 u_{ij}，再根据式（22）修正更新隶属度矩阵 u_{ij}，接着利用式（3）更新 FCM 算法的聚类中心 c_j 。

g) 根据式（25）计算 FCM 目标函数 $J(u,c,M)$，并判断 $\|J(u,c,M)^k-J(u,c,M)^{k-1}\|$ 是否小于 ε 。若 $\|J(u,c,M)^k-J(u,c,M)^{k-1}\|<\varepsilon$，停止计算，输出聚类结果 u_{ij}；否则返回到步骤 f)。

根据上述步骤说明，本文算法的描述实现如下：

输入：蛋白质网络 $G(V,E,P)$ 。

输出：蛋白质复合物。

a)初始化参数：蚂蚁数目为 N，蚁群迭代次数 I，拾起参数 k_p，放下参数 k_d，核心节点阈值 θ，ρ，δ，ω，ε，η；

b)获得蛋白质聚类数目。

(1)for t=1 to I

(2) for n=1 to N

(3) 调用关键蛋白质的选取过程，获得关键蛋白质集合 L

(4) 调用改进的蚁群聚类算法，获得聚类数目 M

(5) end for

(6)end for

c)挖掘蛋白质复合物

(1) for each $v_i\in V$ do

(2) Initialize $FCM\leftarrow L,M$, k=1//初始化 FCM 的初始聚类中心和聚类数目

(3) Compute u_{ij} by using Eq. (2)

(4) Modify u_{ij} by using Eq. (22)

(5) Update c_j by using Eq. (3)

(6) Compute $J(u,c,M)$ by using Eq. (25)

(7)   if $\|J(u,c,M)^k-J(u,c,M)^{k-1}\|<\varepsilon$, STOP

(8)   else

(9)   k=k+1 return to (4)-(7)

(10)   end if

(11) end for

(12)return u_{ij} //得到蛋白质复合物

2.6 算法的时间复杂度

FAC-PC 算法的计算复杂度由以下几个步骤构成：假设 PPI 网络中节点度的最大值为 d_{max}，依据边聚集系数以及基因表达数据构建加权 PPI 网路的时间复杂度为 $O(|E|)$；采用 EPS 公式选取关键蛋白质节点的时间复杂度为 $O(|V|)$；采用节点适应度选取关键组蛋白质的时间复杂度为 $O(|V|^2)$；假设一个节点经过拾起放下完成聚类的比较次数为 $O(INk_3k_1|V|^2)$，符合关键阈值的节点个数为 $k_3|V|$，那么蚁群聚类算法的时间复杂度为 $O(INk_3|V|^2)$；采用 FCM 算法完成最终的复合物识别聚类过程的时间复杂度为 $O(|V|^2)$；因此，FAC-PC 算法的时间复杂度为 $O(|E|+|V|+|V|^2+INk_3|V|^2+|V|^2)$ 即 $O(INk_3|V|^2)$。而在 ACC-FMD 算法中，算法的时间复杂度主要取决于种子节点的选取和蚁群聚类过程，即 $O(NI|V|^3)$；在 ACC-DPC 算法中，算法的时间复杂度主要取决于初始簇的构建和蚁群聚类过程，即 $O(T|V|^3)$；在 GENA 算法中，算法的时间复杂度主要取决于初始化以及优化集群的过程，即 $O(B|V|^3)$；在 WCOACH 算法中，算法的时间复杂度主要取决于初始核的检测和添加附件形成蛋白质复合物的过程，即 $O(\tau V|^3)$；在 IMHRC 算法中，算法的时间复杂度主要取决于主要蛋白质集群形成以及合并修复集群的过程，即 $O(p\gamma\beta|V|^2)$。上述提及的 T、τ、γ、β 和 B 分别表示基因表达时刻数、邻域亲和力阈值、中心获取阈值、中心移除阈值以及预测到的模块数目。

3 实验结果以及分析

3.1 实验环境

FAC-PC 算法实验的编程环境为 Python3.5.2；操作系统

为 Windows 10 家庭中文版；内存 12 GB；处理器为 Intel(R)Core(TM)i5-4200H CPU @ 2.8 GHz。

3.2　实验数据集

为验证本文提出算法的有效性，选用蛋白质相互作用数据相对完整和可靠的酵母蛋白质相互作用网络数据作为实验数据。具体实验数据如下所示：

a）酵母 PPI 网络数据来源于 DIP 数据库[25],去除重复以及自相互作用，该数据库包含 4 995 个蛋白质和 21 554 对相互作用。

b）实验采用的时序基因表达数据为 GSE3431[26],包含 7 079 个蛋白质和 36 个时刻下的基因表达值。

c）本文采用 CYC2008[27]作为标准数据集，该数据集包含 408 个通过生物实验预测得到的蛋白质复合物。

3.3　评价指标

3.3.1 精度、召回率和 F-measure 度量

本文采用文献[28]的精度（*Precision*）、召回率（*Recall*）和 F 度量（*F-measure*）指标来评价算法性能，这些指标的计算依赖于邻域亲和评分。邻域亲和评分主要用来评价预测的复合物与实际复合物的匹配度，其定义为

$$OS(p,b)=\frac{i^2}{|V_p||V_b|} \tag{26}$$

其中：$|V_p|$ 和 $|V_b|$ 分别表示预测复合物 $p=(V_p,E_p)$ 和已知复合物 $b=(V_b,E_b)$ 的规模；i 表示预测复合物和标准复合物交集的规模。若 $OS(p,b)\ge\omega$，则认为 p 和 b 匹配，一般 ω 的取值为 0.2 或者 0.25，本实验中取值为 0.2。设 P 为算法预测结果集合，B 为标准复合物集合，则 P 中至少与一个实际复合物相匹配的复合物数量可表示为 $N_{cp}=|\{p\in P,\exists b\in B,OS(p,b)\ge\omega\}|$；另一方面，$B$ 中至少与一个预测的复合物相匹配的复合物数量为 $N_{cb}=|\{b\in B,\exists p\in P,OS(p,b)\ge\omega\}|$,这复合物检测方法的精度和召回率的定义为

$$precision=\frac{N_{cp}}{|P|} \tag{27}$$

$$recall=\frac{N_{cb}}{|B|} \tag{28}$$

为了避免灵敏度和特异性所带来的偏见,采用 *F-measure* 综合评价指标来评估整体算法的性能。其计算公式如下：

$$F-measure=\frac{2\times precision\times recall}{precision+recall} \tag{29}$$

θ 值度量

随着蛋白质组学研究的深入，使得一个蛋白质与其功能注释向对应成为可能，蛋白质簇发生对于一个给定功能注释在统计学上的意义就可以通过一个超几何分布的等式来进行计算[29]：

$$P-value=1-\sum_{i=0}^{k-1}\frac{\binom{F}{i}\binom{V-F}{C-i}}{\binom{V}{C}} \tag{30}$$

其中：V 代表 PPI 网络中包含的蛋白质总数；C 为预测挖掘出的复合物数目；F 为一个功能组数量；k 为 C 中包含 F 中的蛋白质数目。如果 *P-value* 越小，越接近 0，则说明蛋白质复合物能够随机出现这种功能的概率就越低，可能更具有生物意义。一般地，将 *P-value* 的最小值对应的功能作为该蛋白质复合物的主要功能。通过给每个识别的蛋白质复合物赋予 *P-value* 最小对应的功能，可以预测未知蛋白质的功能。

3.4　参数选择

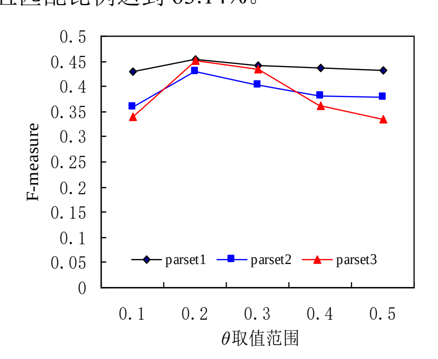
FAC-PC 算法中，由于参数 θ 和 ε 的取值影响实验的聚类效果，所以本文在 15 组 θ 和 ε 的参数取值上独立运行 20 次实验，取 20 次实验的平均值进行分析。实验使用到的参数设置如下：$m=2$，$\rho=0.1$，$I=20$，$\omega=0.2$，$\delta=0.01$ 蚂蚁个数 $N=50$，拾起参数 $k_p=0.9$，放下参数 $k_d=0.2$，$\eta=0.6$。表 1 给出了具体参数设置情况，其中 Set*i* 代表第 *i* 组参数。图 2（a）和（b）中 parset*j* 分别代表 ε 取 0.001 5、0.004 5、0.007 5 对应的 *F-measure* 值和匹配的蛋白质复合物比例，相应的实验结果如图 2 所示。

表 1　实验参数设置

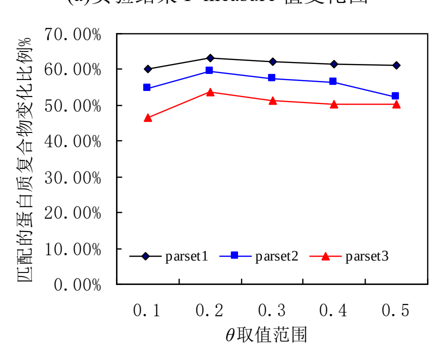
Table 1　Experimental parameter setting

ε 取值范围	θ 取值范围				
	0.1	0.2	0.3	0.4	0.5
0.0015	Set1	Set2	Set3	Set4	Set5
0.0045	Set6	Set7	Set8	Set9	Set10
0.0075	Set11	Set12	Set13	Set14	Set15

由图 2 可知，随着 θ 从 0 到 0.2 逐渐增大，*F-measure* 的值在 ε 不同取值之下也逐渐增大，*F-measure* 达到最大值 0.577，实验识别的复合物和已知的复合物的匹配比例也逐渐增加；随着 θ 从 0.2 到 0.5 逐渐增大，*F-measure* 的值在 ε 不同取值之下逐渐降低，实验识别出的复合物和已知的复合物的匹配比例也逐渐降低。这是因为本文融合边聚集系数与基因表达数据构建加权网络，设计 EPS 公式来选取关键节点，同时利用节点适应度 PFC 度量来逐步遍历关键节点的邻居时，充分考虑内部节点与外部节点之间的联系，随着关键阈值的增大，算法识别的聚类数目逐渐增加，呈上升趋势，实际上每个类中包含的蛋白质数目越少，而类的数目个数就会越多，但是当阈值增大到一定值时，被扩充的节点与关键节点的作用概率要求提高，邻居节点被扩充的可能性就会随之降低，所要求的挖掘的复合物精度逐渐增加，对节点的相关信息要求更高，所以挖掘出的蛋白质复合物会更严格，导致算法 *F-measure* 值和匹配比例先增加后降低。通过观察发现存在一对合理取值即 $\varepsilon=0.0015$，$\theta=0.2$ 使 *F-measure* 达到最大值 0.453 且匹配比例达到 63.14%。



(a)实验结果 F-measure 值变化图



(b)匹配的蛋白质复合物比例变化图

图 2　*F-measure* 值和匹配的蛋白质复合物比例变化图

Fig. 2　*F-measure* and matched protein complex scale change graph

3.5 EPS 度量的有效性分析

为了验证 FAC-PC 算法使用基于基因表达信息和边聚集系数的 EPS 度量公式的有效性，分别基于使用 EPS 度量选取关键蛋白质的 FAC-PC 算法和 ACC-FMD 算法，在 DIP 数据库上进行复合物的识别，实验得到的 *F-measure* 和匹配比例如图 3 所示。

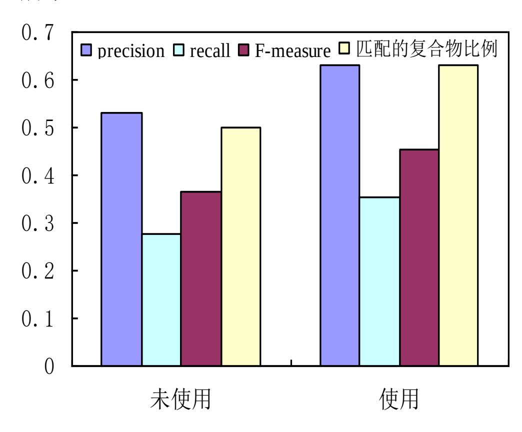


图 3 EPS 度量对比分析

Fig. 3 Comparative analysis of EPS metric

由图 3 显示，使用关键蛋白质选取 EPS 度量的 FAC-PC 算法在 *precision*、*recall*、*F-measure* 取值和匹配的蛋白质复合物比例都比未使用 EPS 度量的取值要高。其中 *precision* 的取值比未使用 EPS 度量提高 19.13%，*recall* 的取值比未使用 EPS 度量提高 27.43%，*F-measure* 的取值比未使用 EPS 度量提高 24.46%，匹配的蛋白质复合物比未使用 EPS 度量提高 26.25%。实验结果说明，使用改进的 EPS 度量的算法的聚类效果得到了提高。这是因为 FAC-PC 充分考虑网络的拓扑特性以及基因共表达程度，根据关键权重度量公式来选择关键蛋白质，进而组成关键组蛋白质来进行聚类。也进一步证明利用关键组蛋白质能很好扩展为一个复合物。

3.6 PFC 和 SI 度量的有效性分析

为了验证 FAC-PC 算法使用 PFC 度量和 SI 度量的有效性，分别基于 PFC 度量以及 SI 度量的 FAC-PC 算法和 ACC-DPC 算法，在 DIP 数据库独立执行 20 次进行复合物的识别，实验检测结果对比分析如图 4 所示。

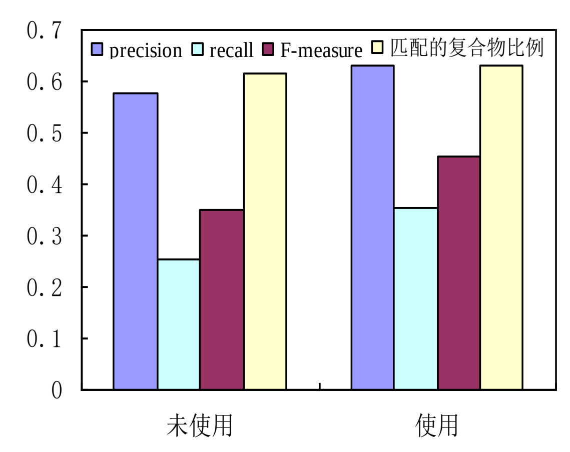


图 4 PFC 和 SI 度量对比分析

Fig. 4 Comparative analysis of PFC and SI metrics

图 4 显示的是使用 PFC 和 SI 度量的 FAC-PC 算法在 *precision*、*recall*、*F-measure* 取值和匹配的蛋白质复合物比例与未使用这两种度量的对比情况，其中使用这两种度量的 *precision* 的取值比未使用这两种度量提高 9.23%，*recall* 的取值比未使用这两种度量提高 39.81%，*F-measure* 的取值比未使用这两种度量提高 28.84%，匹配的蛋白质复合物比未使用这两种度量提高 2.88%。这是因为本文根据 EPS 度量公式选取关键蛋白质节点，同时使用关键组蛋白质代替种子节点进行聚类，考虑到网络的拓扑特性以及基因表达程度，同时也考虑到复合物的模块性以及稠密度，严格控制蚁群拾起放下操作，挖掘的复合物较准确，避免非关键蛋白质无效的拾起放下操作，实验结果说明，使用这两种度量的算法的聚类效果较优。

3.7 算法性能的比较分析

本节将 FAC-PC 分别从精度、召回率和 *F-measure* 的比较分析、聚类效果的比较分析和功能富集的比较分析与 ACC-FMD[13]、ACC-DPC[14]、GENA[17]、WCOACH[18] 和 IMHRC[19]进行比较分析。重复迭代次数 20 次。实验使用到的参数设置如下：$m=2$，$\rho=0.1$，$I=20$，$\delta=0.01$，$\omega=0.2$，蚂蚁个数 $N=50$，拾起参数 $k_p=0.9$，放下参数 $k_d=0.2$，$\varepsilon=0.0015$，$\theta=0.2$，$\eta=0.6$。

1) 精度、召回率和 *F-measure* 的比较分析

为了验证本文算法的性能，将 FAC-PC 算法与其他五种算法在 DIP 数据上独立运行 20 次，取实验结果的平均值进行分析，得到各种算法识别的复合物基本信息以及实验评价指标对比分析如表 2 和图 5 所示。

表 2 各算法识别的复合物的基本信息

Table 2 Basic information of protein complexes for each algorithm

算法	*PM*	*average*	N_{cp}	N_{cb}
ACC-FMD	283	9.5	150	113
ACC-DPC	237	7.8	137	103
GENA	290	5.6	136	87
WCOACH	354	10.3	147	82
IMHRC	366	12.7	210	102
FAC-PC	369	13.5	233	144

在表 2 中，*PM* 表示算法识别出的复合物总数，*average* 是指每个簇中的蛋白质平均个数。由表 2 可以知道，FAC-PC 算法共识别 369 个复合物，每个复合物平均包含 13.5 个蛋白质，其中 233 个预测结果较准确，标准集合中的 144 个复合物可以被算法准确识别到。相较而言，本文算法对于挖掘蛋白质复合物算法具有更高的效率，这是因为本文是基于关键组蛋白质进行蚁群聚类，严格控制蚁群拾起放下操作，同时将通过蚁群聚类得到的聚类数目以及关键蛋白质节点初始化 FCM 算法，在利用改进的隶属度更新策略解决 FCM 的隶属度更新较慢问题，以及综合考虑类内和类间距差异，提出新的目标函数并完成复合物的识别，使得聚类的挖掘效果的准确度和收敛速度加快。

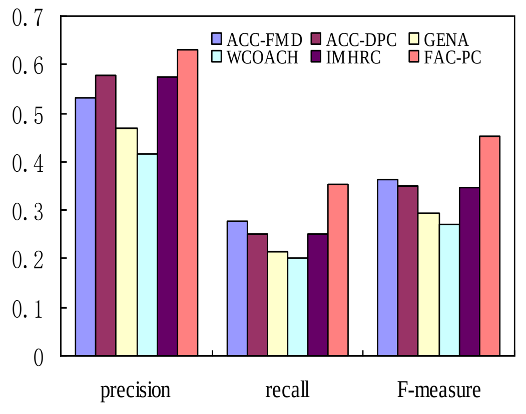


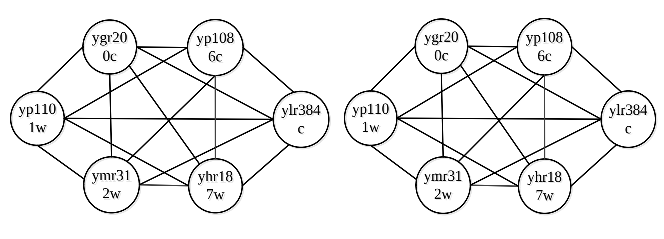
图 5 算法性能比较关系图

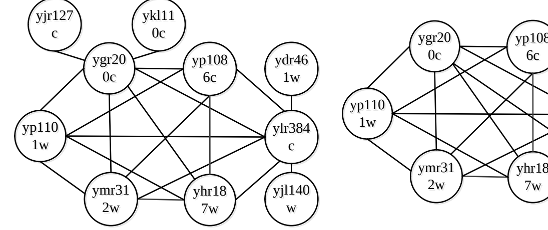
Fig. 5 Performance comparison of algorithm

图 5 显示了各种算法在 DIP 数据集中识别的复合物的结果。从图中可以清晰地发现 FAC-PC 算法在精度、召回率和 *F* 度量指标上取得较好的结果。具体来说，FAC-PC 算法的精度为 63.14%，相较 ACC-FMD、ACC-DPC、GENA、WCOACH 和 IMHRC 分别提高了 19.13%、9.23%、34.64%、52.06%和 10.05%；召回率为 35.29%，相较 ACC-FMD、ACC-DPC、GENA、WCOACH 和 IMHRC 分别提高了 27.43%、39.81%、65.51%、75.61%和 41.18%；F 值度量为 45.28%，相较 ACC-FMD、ACC-DPC、GENA、WCOACH 和 IMHRC 分别提高了 24.46%、28.84%、54.48%、67.17%和 30.02%。实验结果表明，使用本文算法挖掘蛋白质复合物的聚类精度、召回率和 *F-measure* 相比较其他五种算法都得到了提高。这是因为，在 ACC-FMD 算法中，使用种子节点进行蚁群聚类，若种子节点之间的相似度较大，会出现重复的拾起放下操作，会挖掘出重叠模块，且存在大量的合并过滤，导致挖掘的时间效率低；在 ACC-DPC 算法中，初始聚类中心的选取仅仅考虑到网络的拓扑特性聚类系数，没有考虑到蛋白质基因生物信息，选择簇中心的条件比较单一，仅仅根据拾起放下规则，在聚类的过程中存在大量的拾起放下操作，导致挖掘出的效果不佳；在 GENA 算法中，使用贪婪方法初始化集群，在聚类系数的基础上选取种子节点，仅仅考虑了网络的拓扑特性，挖掘的效果存在大量的重叠模块；在 WCOACH 算法中，仅仅利用 GO 信息来构建加权网络，缺乏考虑蛋白质网络本身的拓扑特性以及特征，且在聚类时，若核心节点选取较为相似，则会挖掘出大量重叠的模块，最终导致挖掘的准确性降低；在 IMHRC 算法中，构建加权 PPI 网络时，仅仅考虑节点度即网络的拓扑结构，没有融合生物信息，考虑构建的加权 PPI 网络比较单一，使得挖掘聚类效果不佳。而本文是综合考虑网络的拓扑结构和生物基因表达信息来构建加权网络，基于关键组蛋白质进行蚁群聚类，同时根据适应度来严格控制拾起放下操作，在最终使用 FCM 算法完成聚类的时候，综合考虑类内和类间距以及设计隶属度更新策略来改进隶属度更新较慢等问题，可以较为精确和快速地挖掘出蛋白质复合物。因此，本文提出的算法的聚类效果较好。

2) 聚类效果的比较分析

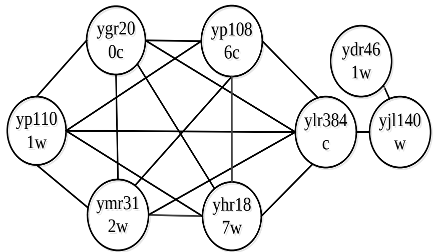
为评估本文提出的 FAC-PC 算法的聚类效果，将本文算法与其他五种算法挖掘的 Elongator holoenzyme 复合物可视化进行对比分析聚类效果，聚类结果如图 6 所示。图 6 显示了不同算法检测到的 Elongator holoenzyme 复合物结果，其中图 6(a)是该标准复合物所包含的蛋白质相互作用情况；(b)是本文算法的检测结果；(c)是 ACC-FMD 算法的检测结果；(d)是 ACC-DPC 算法的检测结果；(e)是 GENA 算法的检测结果；(f)是算法 WCOACH 的检测结果；(g)是 IMHRC 算法的检测结果。通过图 6 显示，本文算法能够准确地识别蛋白质复合物；ACC-FMD 算法识别出标准复合物中的 6 个蛋白质，但是也包含了 4 个非 Elongator holoenzyme 复合物内的蛋白质；ACC-DPC 算法识别出标准复合物中的 6 个蛋白质，但是也包含了 1 个非 Elongator holoenzyme 复合物内的蛋白质；GENA 算法识别出标准复合物中的 6 个蛋白质，但是也包含了 2 个非 Elongator holoenzyme 复合物内的蛋白质；WCOACH 算法识别出标准复合物中的 5 个蛋白质；IMHRC 算法识别出标准复合物中的 6 个蛋白质，但是也包含了 3 个非 Elongator holoenzyme 复合物内的蛋白质。实验结果表明，本文算法挖掘的蛋白质复合物聚类效果较好。这是因为，本文通过蛋白质网络的拓扑特性和基因表达信息来构建加权网络，可以降低假阳性的影响；同时根据 EPS 度量选取关键蛋白质，在通过节点适应度 PFC 度量来进一步选取关键组蛋白质，利用关键组蛋白质进行蚁群聚类，这样减少大量的拾起放下和重复的合并过滤操作，进而提高聚类运行效率和准确性；将得到的关键蛋白质节点以及聚类数目初始化 FCM，接着根据隶属度更新的改进策略改进隶属度更新较慢的缺陷，以及综合考虑类内和类间距，优化 FCM 的目标函数；最后利用改进的 FCM 完成蛋白质复合物的挖掘。实验结果表明，本文算法在识别蛋白质复合物上具有较好的聚类效果。

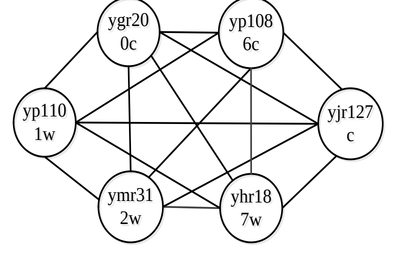




(a) 标准复合物 (b)FAC-PC 算法检测的复合物 (c)ACC-FMD 算法检测的复合物 (d)ACC-DPC 算法检测的复合物

(a) Standard complexes (b) Compounds detected by FAC-PC algorithm (c) Complexes detected by ACC-FMD algorithm (d) Complexes detected by ACC-DPC algorithm





(e)GENA 算法检测的复合物 (f)WCOACH 算法检测的复合物 (g)IMHRC 算法检测的复合物

(e) Complexes detected by GENA algorithm (f) Complexes detected by WCOACH algorithm (g) Complexes detected by IMHRC algorithm

图 6 各个算法的复合物挖掘可视化比较

Fig. 6 Visualization comparison of protein complexes of each algorithm

3) 功能富集的比较分析

为了测试算法识别的复合物的生物学意义，本文采用功能富集分析。*P-value* 被认为是衡量识别的复合物是一个真正的蛋白质复合物的可能性。识别的复合物的低 *P-value* 值表

明该复合物具有很高的统计学意义，一般将 *P-value* 的最小值对应的功能作为该功能模块的主要功能，通过给每个识别的复合物赋予最小的 *P-value* 值对应的功能，可以识别预测复合物的功能。若一个模块的 *P-value*<0.01，则认为这个复合物是显著的。显著的复合物数量在识别出的复合物总数中所占的比例可以很好地评价各个算法的整体性。具体各个算法性能比较分析如表 3 所示。

表 3　各个算法识别的复合物的显著性统计信息

Table 3　Significant statistics of protein complexes detected by each algorithm

算法	*PM*	*SC*	Proportion
ACC-FMD	283	141	49.82%
ACC-DPC	237	160	67.51%
GENA	290	135	46.55%
WCOACH	354	263	74.29%
IMHRC	366	180	49.18%
FAC-PC	369	305	82.66%

在表 3 中，*PM* 表示算法识别出的复合物总数，*SC* 是具有显著意义的复合物数目。FAC-PC 算法识别的复合物数目中显著性复合物的比例达到 82.66%，相比较 ACC-FMD、ACC-DPC、GENA、WCOACH 和 IMHRC 算法分别提高了 65.92%、22.44%、77.57%、11.27%、68.08%。由此可见，FAC-PC 算法识别出的复合物具有很强的生物统计学意义。这是因为本文提出的算法在构建加权网络的时候，综合考虑网络的拓扑特性和基因共表达程度，同时利用关键组蛋白质来进行蚁群聚类，将关键蛋白质节点以及聚类数目初始化 FCM，根据隶属度相对更新策略来改进隶属度的更新较慢的问题，同时还综合考虑类内和类间距对实验结果产生的影响，提出新的目标函数，最终导致聚类效果较好，执行效率高，挖掘的生物蛋白质复合物更具有生物统计意义。

表 4 具体给出本文 FAC-PC 算法识别出的复合物实例。其中 *OS* 表示复合物的匹配率，*SM* 表示的是正确匹配的蛋白质个数，Predicted protein 表示组成复合物的所有蛋白质，加粗部分表示被匹配的蛋白质。从表 4 可以看出，当 *P-value*=2.22E-18 时，本文算法识别的 NatC 复合物的匹配率达到了 0.82，正确匹配的蛋白质个数是 9，这是因为 YGR134W 和 YNL288W 蛋白质与复合物内部连接比较松散。由此可见，FAC-PC 算法识别的蛋白质复合物效果更好。

表 4　FAC-PC 算法识别的复合物实例

Table 4　Instances of protein complexes detected by FAC-PC algorithm

P-value	OS	SM	Predicted protein	Real complex
2.22E-18	0.82	9	**YCR093W YER068W** YGR134W **YPR072W** YNL288W **YIL038C YNR052C YDL165W YAL021C YDR443C YPL011C**	NatC
2.18E-29	0.79	20	**YLR421C YDL007W YER021W** YPR108W YHR200W **YDL097C YKL145W** YGR232W **YFR004W YDL147W YOR261C YDR427W YOR259C YOR117W YGL048C YHR027C YIL075C YFR010W YFR052W YDR394W YGL004C YBR272C YER012W**	Giplp /Gle7p
1.90E-32	0.65	12	**Y0R116C YNL151C YJL011C YKR025W** YDL150W **YNR003C YDR045C YKL144C YOR224C YPR190C YOR207C YPR110C YDR005C**	Cde28p/Clb5p
1.56E-29	0.88	23	**YBR253W YPR070W YMR112C YGR104C YOL051W YCR081W YOL135C YGL152W YDL005C YBR193C YDR308C YLR071C YGL025C YPL042C** YHR058C **YNR010W YHR042C YNL025C YOR174W YPR168W YER022W YNL236W YBL093C YPL248C**	Glycine cleavage

4　结束语

本文在结合边聚集系数与基因表达数据构建的加权蛋白质网络，提出一种基于模糊蚁群聚类的蛋白质复合物识别算法 FAC-PC。基于边聚集系数与基因表达数据，设计 EPS 公式选取关键蛋白质节点，基于期望稠密度与模块度改进节点适应度，提出 PFC 公式获取关键组蛋白质；基于权重改进相似度计算，设计 SI 相似度度量优化蚁群算法的拾起放下概率，利用关键组蛋白质完成蚁群聚类，获得聚类数目；最后将关键节点以及聚类数目初始化 FCM 算法，同时改进隶属度的更新公式，提出兼顾类内和类间距新的 FCM 目标函数，最终完成复合物的识别。为了评估算法的性能，本文将 FAC-PC 算法与其他五种算法进行了对比。实验结果表明，FAC-PC 算法具有更高的准确率、召回率，识别的复合物具有更强的生物统计意义。今后可以将 FAC-PC 算法应用于疾病预测和关键蛋白质识别等相关研究中。

参考文献：

[1] 冀俊忠，高光轩. 基于文化算法的 PPI 网络功能模块检测方法 [J]. 北京工业大学学报，2017，43 (1)：0013-0021. (Ji Junzhong, Gao Guangxuan. Detecting functional module method based on cultural algorithm in protein-protein interaction networks [J]. Journal of Beijing University of Technology, 2017, 43 (1): 0013-0021.)

[2] 郑文萍，李晋玉，王杰. 基于遗传算法的蛋白质复合物识别算法 [J]. 计算机科学与探索，2018，12 (5)：794-803. (Zheng Wenping, Li Jinyu, Wang Jie. Protein complex recognition algorithm based on genetic algorithm [J]. Journal of Frontiers of Computer Science and Technology, 2018, 12 (5): 794-803.)

[3] Cai Leixin, Chen Rong, Lyu Qiang. Selecting the best of the predictions of compound structures for protein docking by spectral clustering [J]. Journal of Chinese Computer Systems, 2015, 36 (10): 2365-2368.

[4] 李敏，王建新，刘彬彬，等. 基于极大团扩展的蛋白质复合物识别算法 [J]. 中南大学学报，2010，41 (2)：560-565. (Li Min, Wang Jianxin, Liu Bin bin, *et al.* An algorithm for identifying protein complexes based

on maximal clique extension [J]. Journal of Central South University, 2010, 41 (2): 560-565.)

[5] Kessler J, Andrushchenko V, Kapitan J, *et al.* Insight into vibrational circular dichroism of proteins by density functional modeling [J]. Physical Chemistry Chemical Physics, 2018, 20 (7): 4926-4935.

[6] Aldeco R, Marin I. Jerarca: efficient analysis of complex networks using hierarchical clustering [J]. PLoS ONE, 2010, 5 (7): 11585-11591.

[7] Abeysirigunawardena S C, Kim H, Lai J, *et al.* Evolution of protein-coupled RNA dynamics during hierarchical assembly of ribosomal complexes [J]. Nature Communications, 2017, 8 (1): 492-500.

[8] 雷秀娟, 高银, 郭玲. 基于拓扑势加权的动态 PPI 网络复合物挖掘方法 [J]. 电子学报, 2018, 46 (1): 145-151. (Lei Xiujuan, Gao Yin, Guo Ling. Ming protein complexes based on topology potential weight in dynamic protein-protein interaction networks [J]. Acta Electronica Sinica, 2018, 46 (1): 145-151.)

[9] Yao Xiaohui, Yan Jingwen, Liu Kefei, *et al.* Tissue-specific network-based genome wide study of amygdala imaging phenotypes to identify functional interaction modules [J]. Bioinformatics, 2017, 33 (20): 3250-3257.

[10] Trivodaliev K, Cingovska I, Kalajdziski S. Protein function prediction by spectral clustering of protein interaction network [J]. Communications in Computer & Information Science, 2011, 8 (25): 108-117.

[11] 冀俊忠, 杨明浩, 杨翠翠, 等. 快速的基于蚁群聚类的 PPI 网络功能模块检测方法 [J]. 北京工业大学学报, 2016, 42 (8): 1182-1192. (Ji Junzhong, Yang Minghao, Yang Cuicui, *et al.* Fast ant colony clustering for functional module detection algorithm in PPI networks [J]. Journal of Beijing University of Technology, 2016, 42 (8): 1182-1192.)

[12] Ji Junzhong, Liu Hongxin, Zhang Aidong, *et al.* ACC-FMD: ant colony clustering for functional module detection in protein-protein interaction networks [J]. International Journal of Data Mining & Bioinformatics, 2015, 11 (3): 331-363.

[13] 赵学武, 程新党, 吕嘉伟, 等. 融合时序保持特征和蚁群聚类的动态 PPI 网络复合物识别 [J]. 小型微型计算机系统, 2017, 38 (6): 1311-1316. (Zhao Xuewu, Cheng Xindang, Lyu Jiawei, *et al.* Identify protein complexes by integrating temporal function continue feature and ant colony clustering on dynamic PPI networks [J]. Journal of Chinese Computer Systems, 2017, 38 (6): 1311-1316.)

[14] Lei Xiujuan, Wu Fangxiang, Tian Jianfang, *et al.* ABC and IFC: modules detection method for PPI network [J]. Biomed Research International, 2014, 8 (1): 968173-968183.

[15] 张媛, 贾克斌, Zhang Aidong. 基于多视图融合的蛋白质功能模块检测方法 [J]. 电子学报, 2014, 42 (12): 2337-2344. (Zhang Yuan, Jia Kebin, Zhang Aidong. Consistent protein functional module detection from multi-view of biological data [J]. Acta Electronica Sinica, 2014, 42 (12): 2337-2344.)

[16] Dimitrakopoulos C, Theofilatos K, Pegkas A, *et al.* Predicting overlapping protein complexes from weighted protein interaction graphs by gradually expanding dense neighborhoods [J]. Artificial Intelligence in Medicine, 2016, 71 (7): 62-69.

[17] Kouhsar M, Zaremirakabad F, Jamali Y. WCOACH: Protein complex prediction in weighted PPI networks [J]. Genes & Genetic Systems, 2016, 91 (1): 47-47.

[18] Ama M, Eslahchi C. Discovering overlapped protein complexes from weighted PPI networks by removing inter-module hubs [J]. Scientific Reports, 2017, 7 (1): 3247-3260.

[19] Kesemen O, Tezel O, Ozkul E. Fuzzy c-means clustering algorithm for directional data (FCM4DD) [J]. Expert Systems with Applications, 2016, 58 (C): 76-82.

[20] 倪问尹, 王建新, 熊慧军, 等. 基于不确定数据的功能模块预测 [J]. 四川大学学报, 2013, 45 (5): 0080-0087. (Ni Wenyin, Wang Jianxin, Xiong Huijun, *et al.* Research of detecting functional modules based on uncertainty data [J]. Journal of Sichuan University, 2013, 45 (5): 0080-0087.)

[21] 李敏, 张含会, 费耀平. 融合 PPI 和基因表达数据的关键蛋白质识别方法 [J]. 中南大学学报, 2013, 44 (3): 1024-1029. (Li Min, Zhang Hanhui, Fei Yaoping. Essential protein discovery method based on integration of PPI and gene expression data [J]. Journal of Central South University, 2013, 44 (3): 1024-1039.)

[22] 赵碧海, 李学勇, 胡赛, 等. 基于关键功能模块挖掘的蛋白质功能预测 [J]. 自动化学报, 2018, 44 (1): 183-192. (Zhao Bihai, Li Xueyong, Hu Sai, *et al.* Prediction of protein functions based on essential functional modules mining [J]. Acta Automatica Sinica, 2018, 44 (1): 183-192.)

[23] 刘翠翠, 孙伟. 基于加权网络和局部适应度的蛋白质复合物识别算法 [J]. 计算机应用研究, 2018, 35 (8): 2308-2310. (Liu Cuicui, Sun Wei. Algorithm for identifying protein complexes based on weighted network and local fitness [J]. Application Research of Computers, 2018, 35 (8): 2308-2310.)

[24] Xie X L, Beni G. A validity measure for fuzzy clustering [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 1991, 13 (8): 841-847.

[25] Li Xiaoli, Wu Min, Kwoh C K, *et al.* Computational approaches for detecting protein complexes from protein interaction networks: a survey [J]. BMC Genomics, 2010, 11 (Suppl 1): 1-19.

[26] Tu B P, Kudlicki A, Rowicka M, *et al.* Logic of the yeast metabolic cycle: temporal compartmentalization of cellular processes [J]. Science, 2005, 310 (5751): 1152-1158.

[27] Pu S, Wong J, Turner B, *et al.* Up-to-date catalogues of yeast protein complexes [J]. Nucleic Acids Research, 2009, 37 (3): 825-831.

[28] Zhang Yijia, Lin Hongfei, Yang Zhihao, *et al.* A method for predicting protein complex in dynamic PPI networks [J]. BMC Bioinformatics, 2016, 17 (Suppl 7): 229-239.

[29] Lei Xiujuan, Wu Shuang, Liang Ge, *et al.* Clustering and overlapping modules detection in PPI network based on IBFO [J]. Proteomics, 2013, 13 (2): 278-290.